



# Will It Fit? The Local AI VRAM Cheat Sheet

Quick answers for running Ollama models on your own PC. From FitGPU, fitgpu.com

## How much VRAM does each model need?

| Ollama model      | Params | Q4_K_M   | Q8_0     | 8 GB    | 12 GB   | 16 GB   | 24 GB   | Mini-PC: give GPU |
|-------------------|--------|----------|----------|---------|---------|---------|---------|-------------------|
| llama3.1:8b       | 8.03B  | ~6.4 GB  | ~10.4 GB | Fits    | Fits    | Fits    | Fits    | 7+ GB             |
| mistral:7b        | 7.25B  | ~5.9 GB  | ~9.5 GB  | Fits    | Fits    | Fits    | Fits    | 7+ GB             |
| qwen3:8b          | 8.19B  | ~6.6 GB  | ~10.7 GB | Fits    | Fits    | Fits    | Fits    | 7+ GB             |
| gemma2:9b         | 9.24B  | ~8.1 GB  | ~12.7 GB | Offload | Fits    | Fits    | Fits    | 9+ GB             |
| qwen3:14b         | 14.8B  | ~11.0 GB | ~18.5 GB | Offload | Tight   | Fits    | Fits    | 12+ GB            |
| deepseek-r1:14b   | 14.8B  | ~11.2 GB | ~18.6 GB | Offload | Tight   | Fits    | Fits    | 12+ GB            |
| mistral-small:24b | 23.6B  | ~16.9 GB | ~28.8 GB | Offload | Offload | Offload | Fits    | 19+ GB            |
| gemma2:27b        | 27.2B  | ~20.2 GB | ~33.8 GB | Offload | Offload | Offload | Fits    | 22+ GB            |
| qwen3:30b-a3b     | 30.5B  | ~21.2 GB | ~36.5 GB | Offload | Offload | Offload | Fits    | 24+ GB            |
| qwen3:32b         | 32.8B  | ~23.4 GB | ~39.9 GB | Offload | Offload | Offload | Tight   | 26+ GB            |
| llama3.3:70b      | 70.6B  | ~48.9 GB | ~84.4 GB | Offload | Offload | Offload | Offload | 54+ GB            |

**These are estimates, not measurements.** Same formula as the FitGPU picker: weights (params × bits per weight) + KV cache at 4K context (F16) + 0.5 GB + 10% of weights. Fit columns use Q4\_K\_M. "Fits" = under 90% of VRAM, "Tight" = 90-100%, "Offload" = part of the model spills to system RAM and slows down a lot. "Mini-PC" = shared memory (e.g. Ryzen AI / Radeon 890M) you'd need to assign to the GPU for a comfortable fit. Real Ollama downloads can differ slightly.

## Which VRAM tier runs what (Q4\_K\_M, 4K context)

| Tier                                              | Comfortable picks                                                                                                                       |
|---------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------|
| <b>8 GB</b>                                       | 7-9B models: llama3.1:8b, mistral:7b, qwen3:8b (gemma2:9b needs 12 GB)                                                                  |
| <b>12 GB</b> (e.g. RTX 3060 12 GB)                | All 7-9B models with room for longer context; 14B is tight                                                                              |
| <b>16 GB</b> (RX 7600 XT, RX 6950 XT, RX 9070 XT) | 14B models (qwen3:14b, deepseek-r1:14b) fit; 24B+ needs offload                                                                         |
| <b>24 GB</b> (RTX 3090, RX 7900 XTX, RTX 4090)    | Up to mistral-small:24b, gemma2:27b, qwen3:30b-a3b; 32B is tight; 70B needs offload                                                     |
| <b>Mini-PC shared memory</b>                      | Lots of capacity, lower bandwidth: big MoE models (qwen3:30b-a3b, gpt-oss:20b) run fine; dense models are slower (see measured results) |

## Real measured speeds from FitGPU

| Hardware                                                         | Model                       | Quant  | Gen tok/s                               | Memory used |
|------------------------------------------------------------------|-----------------------------|--------|-----------------------------------------|-------------|
| RX 6950 XT 16 GB (Ryzen 9 7900X desktop)                         | llama3.1:8b                 | Q4_K_M | <b>99.4</b>                             | 5.3 GB      |
| RX 6950 XT 16 GB (Ryzen 9 7900X desktop)                         | mistral:7b                  | Q4_K_M | <b>105.9</b>                            | 4.9 GB      |
| RX 6950 XT 16 GB (Ryzen 9 7900X desktop)                         | phi4:14b                    | Q4_K_M | <b>55.6</b>                             | 9.7 GB      |
| RX 6950 XT 16 GB (Ryzen 9 7900X desktop)                         | gpt-oss:20b                 | MXFP4  | <b>115.8</b>                            | 12.7 GB     |
| RX 6950 XT 16 GB (Ryzen 9 7900X desktop)                         | qwen3.5:27b                 | Q4_K_M | <b>9.9</b> (81% on GPU, rest offloaded) | 14.2 GB     |
| Radeon 890M, shared memory (MinisForum AI X1 Pro-470, 64 GB RAM) | qwen3:30b-a3b instruct-2507 | Q4_K_M | <b>29.2</b>                             | 18.9 GB     |
| Radeon 890M, shared memory (MinisForum AI X1 Pro-470, 64 GB RAM) | gpt-oss:20b                 | MXFP4  | <b>18.2</b>                             | 12.7 GB     |

| Hardware                                                         | Model           | Quant  | Gen tok/s   | Memory used |
|------------------------------------------------------------------|-----------------|--------|-------------|-------------|
| Radeon 890M, shared memory (MinisForum AI X1 Pro-470, 64 GB RAM) | llama3.1:8b     | Q4_K_M | <b>14.0</b> | 5.3 GB      |
| Radeon 890M, shared memory (MinisForum AI X1 Pro-470, 64 GB RAM) | gemma3:12b      | Q4_K_M | <b>8.1</b>  | 8.1 GB      |
| Radeon 890M, shared memory (MinisForum AI X1 Pro-470, 64 GB RAM) | qwen3:14b       | Q4_K_M | <b>7.6</b>  | 9.7 GB      |
| Radeon 890M, shared memory (MinisForum AI X1 Pro-470, 64 GB RAM) | deepseek-r1:14b | Q4_K_M | <b>7.8</b>  | 9.5 GB      |

Measured by FitGPU on our own machines with Ollama 0.40.4 on Windows 11, Oct 15-17, 2026: median of 3 runs, 4K context, 256 generated tokens. Full list: [fitgpu.com/benchmarks.html](https://fitgpu.com/benchmarks.html)

5 quick tips

**Q4\_K\_M is the sweet spot.** About 4.85 bits per weight vs 8.5 for Q8\_0, so a model takes roughly 43% less space with only a small quality drop.

**Context length eats VRAM.** The KV cache grows with context. llama3.1:8b uses ~0.5 GB of KV cache at 4K but ~4.3 GB at 32K (estimate, F16).

**Offloading works, but it's slow.** qwen3.5:27b with 81% on GPU ran at 9.9 tok/s on a 16 GB card, vs ~100 tok/s for 8B models that fully fit.

**MoE models punch above their size.** qwen3:30b-a3b only activates ~3B params per token: 29.2 tok/s on a mini-PC iGPU, about 4x a dense 14B model on the same machine.

**Measure, don't guess.** Run the free FitGPU bench kit before and after an upgrade to see the real difference.

**Find what fits your PC: [fitgpu.com](https://fitgpu.com)**  
Free model-fit picker, real benchmarks, and a free bench kit ([fitgpu.com/submit.html](https://fitgpu.com/submit.html))

Questions?  
[lcp@lcp-solutionsllc.com](mailto:lcp@lcp-solutionsllc.com)

© 2026 FitGPU. Estimates are labeled as estimates; formulas at [fitgpu.com/methodology.html](https://fitgpu.com/methodology.html). GPU VRAM specs from vendor/Wikipedia spec tables.